Probabilistic Machine Learning Estimation of Ocean Mixed Layer Depth from Dense Satellite and Sparse In-Situ Observations

Dallas Foster[1], David John Gagne[2], and Daniel Bridger Whitt[3]

[1]Oregon State University
[2]NCAR
[3]National Center for Atmospheric Research

November 23, 2022

Abstract

The ocean mixed layer plays an important role in subseasonal climate dynamics because it can exchange large amounts of heat with the atmosphere, and it evolves significantly on subseasonal timescales. Estimation of the subseasonal variability of the ocean mixed layer is therefore important for subseasonal to seasonal prediction and analysis. The increasing coverage of in-situ Argo ocean profile data allows for greater analysis of the aseasonal ocean mixed layer depth (MLD) variability on subseasonal and interannual timescales; however, current sampling rates are not yet sufficient to fully resolve subseasonal MLD variability. Other products, including gridded MLD estimates, require optimal interpolation, a process that often ignores information from other oceanic variables. We demonstrate how satellite observations of sea surface temperature, salinity, and height facilitate MLD estimation in a pilot study of two regions: the mid-latitude southern Indian and the eastern equatorial Pacific Oceans. We construct multiple machine learning architectures to produce weekly 1/2 degree gridded MLD anomaly fields (relative to a monthly climatology) with uncertainty estimates. We test multiple traditional and probabilistic machine learning techniques to compare both accuracy and probabilistic calibration. We find that incorporating sea surface data through a machine learning model improves the performance of MLD estimation over traditional optimal interpolation in terms of both mean prediction error and uncertainty calibration. These preliminary results provide a promising first step to greater understanding of aseasonal MLD phenomena and the relationship between the MLD and sea surface variables. Extensions to this work include global and temporal analyses of MLD.

Probabilistic Machine Learning Estimation of Ocean Mixed Layer Depth from Dense Satellite and Sparse In-Situ Observations

Dallas Foster[1], David John Gagne II[2], Daniel B. Whitt[2]

[1]Department of Mathematics, Oregon State University, Corvallis, OR 97331
[2]National Center for Atmospheric Research, Boulder, CO, 80305

Key Points:

- Machine learning models that incorporate surface and ocean profile data improve ocean MLD estimates.
- Model performance is dependent on spatial location and strength of the sub-seasonal variance.
- Probabilistic sampling techniques capture uncertainty better than standard or parametric approaches.

Corresponding author: David John Gagne II, `dgagne@ucar.edu`

Abstract

The ocean mixed layer plays an important role in subseasonal climate dynamics because it can exchange large amounts of heat with the atmosphere, and it evolves significantly on subseasonal timescales. Estimation of the subseasonal variability of the ocean mixed layer is therefore important for subseasonal to seasonal prediction and analysis. The increasing coverage of in-situ Argo ocean profile data allows for greater analysis of the aseasonal ocean mixed layer depth (MLD) variability on subseasonal and interannual timescales; however, current sampling rates are not yet sufficient to fully resolve subseasonal MLD variability. Other products, including gridded MLD estimates, require optimal interpolation, a process that often ignores information from other oceanic variables. We demonstrate how satellite observations of sea surface temperature, salinity, and height facilitate MLD estimation in a pilot study of two regions: the mid-latitude southern Indian and the eastern equatorial Pacific Oceans. We construct multiple machine learning architectures to produce weekly 1/2 degree gridded MLD anomaly fields (relative to a monthly climatology) with uncertainty estimates. We test multiple traditional and probabilistic machine learning techniques to compare both accuracy and probabilistic calibration. We find that incorporating sea surface data through a machine learning model improves the performance of MLD estimation over traditional optimal interpolation in terms of both mean prediction error and uncertainty calibration. These preliminary results provide a promising first step to greater understanding of aseasonal MLD phenomena and the relationship between the MLD and sea surface variables. Extensions to this work include global and temporal analyses of MLD.

Plain Language Summary

The top layer of the ocean, called the surface mixed layer, features temperature and salinity that are relatively uniform throughout its depth. The depth of this layer can vary depending on the exact location, time of year and is impacted by many physical processes. Although it is typically only a few percent of the ocean depth, the mixed layer is important because it regulates heat exchange between the deep ocean and the atmosphere, and it hosts virtually all photosynthesis that sustains ocean ecosystems. Observations of the mixed layer depth are infrequent in time and space because of the size of the ocean in comparison to the number of observing instruments. Satellite data is widely available for information about the surface of the ocean, but unfortunately there is not an exact relationship between the surface information and the mixed layer depth. In this paper, we study machine learning models' abilities to learn this relationship with the available data and to produce reasonable fine-scale estimates of the mixed layer depth. In particular, we emphasize the ability of the machine learning model to estimate how uncertain it is about its estimates.

1 Introduction

Because of the ocean surface mixed layer's role as intermediary between ocean and atmosphere, many important processes, such as water mass formation and ocean circulation (Hanawa & Talley, 2001; Stommel, 1979) and air-sea interaction (Frankignoul & Hasselmann, 1977; Kraus & Turner, 1967) are sensitive to the ocean surface mixed layer depth (MLD). While there have been several recent efforts to observe and quantify the global climatological behavior of the MLD based on the in-situ array of thousands of vertically-profiling Argo floats (Holte et al., 2017; Schmidtko et al., 2013; D. B. Whitt et al., 2019), little effort has been devoted to quantifying the subseasonal and interannual (aseasonal) variability of the MLD because the Argo array is not sufficiently large to fully resolve subseasonal MLD variability. Through this study, we take a preliminary step toward improved observational estimates of aseasonal MLD variability by investigating the relationship between MLD and sea surface salinity, temperature, and height anomalies.

Due largely to the increasing coverage of the Argo array (Holte et al., 2017), the MLD is increasingly well-observed globally. Despite this improvement, however, the data is insufficient to recover sub-seasonal processes on a fine grid at high frequency. Modern attempts to recover variables using a hybrid data collection of in-situ and satellite data typically use optimal interpolation (Roemmich & Gilson, 2009; Guinehut et al., 2012). Our aim in this paper is to demonstrate the utility of informing MLD estimation using satellite surface data through a machine learning framework.

The application of machine learning to the geosciences is a rapidly growing field ((Monteleoni et al., 2013; Reichstein et al., 2019; Weyn et al., 2019; Lary et al., 2016; Irrgang et al., 2020). The machine learning approach offers a flexible, data-driven route to regression and classification tasks that has been used for parameterizations (Bolton & Zanna, 2019; Gagne et al., 2020; Rasp et al., 2018; O'Gorman & Dwyer, 2018; Gentine et al., 2018; Jiang et al., 2018; Brenowitz & Bretherton, 2018), forecasting (Pathak et al., 2018; McGovern et al., 2017; Ukkonen & Mäkelä, 2019; Irrgang et al., 2020; Weyn et al., 2019; Hsieh & Tang, 1998), data assimilation (R. Cintra et al., 2016; Wahle et al., 2015; R. S. Cintra & Velho, 2018), and remote sensing (Lary et al., 2016; Ouali et al., 2017). The commonality to many of these approaches and the motivation for use in this study is not only the lack of a deterministic model between the sea surface variables and the mixed layer depth, but also the possibility of an empirical model being learned from the existing data. Unfortunately, many successes in machine learning research are also in over-determined regimes, in which the amount of data is large in comparison to the number of independent parameters. Extrapolation regimes, where data are sparse in one or more dimensions, are known to be problematic because the prediction depends more heavily on the underlying assumptions of the model. This is particularly problematic in oceanography, where many unknown quantities are 2 or 3 dimensional, and data availability is still relatively sparse.

While the study of machine learning can trace its history to Rosenblatt's perceptron (Rosenblatt, 1958), the implementation of early machine learning methods and architectures in a data-driven way was considered computationally infeasible for moderate to large applications until the late 1980s with the development of the back-propagation algorithm (Rumelhart et al., 1986), which enabled training of multi-layered neural networks. Despite advances through the nineties and early twenty-first century, the deep learning revolution did not occur until 2006 (Goodfellow et al., 2016) when an explosion of reliable training data, computing power, neural network layers, and regularization techniques have dramatically increased neural network accuracy. As demonstrated in Guo et al. (2017), this improvement in accuracy has also hindered the capacity of neural networks to be well-calibrated, i.e. when forecast probabilities match the system's true probabilities, and hence offer accurate representations of the underlying probability distributions. The ability for a neural network to be well-calibrated is of critical importance. Data Assimilation research has repeatedly shown that proper estimation of the background error covariance can improve reconstruction estimates (Valler et al., 2019). In the estimation of sea surface temperature or sea level anomaly, mis-quantification of atmospheric uncertainties has also been shown to cause significant and non-local errors in reanalysis estimates(Chaudhuri et al., 2016). Parallel developments have led to the field of probabilistic neural networks to address this calibration problem in machine learning.

The ultimate goal of probabilistic neural networks is to be able to accurately and precisely define the posterior probability distribution conditioned on the data. Using a Bayesian framework allows us to easily account for sources of error and randomness in the data, weights, or model. The gold standard for this task is often sampling from the posterior distribution using a Markov Chain Monte Carlo (MCMC) scheme (Brooks, 2011; Gelman et al., 2013), but this approach is still computationally infeasible for modern neural networks. There have been several approximations and techniques developed for producing estimates of the posterior probability including the development of Bayesian Neural Networks, with weight uncertainty (Neal, 1996; Blundell et al., 2015), Stochastic Gradient Langevin Dynamics (Welling & Teh, 2011), Variational Inference (Paisley et al., 2012; Hoffman & Blei, 2015; Kingma et al., 2015), Probabilistic Backpropagation (Rezende et al., 2014; Hernández-Lobato & Adams, 2015), Dropout (Hinton et al., 2012; Ba & Frey, 2013; Maeda, 2014; Gal & Ghahramani, 2016; Gal et al., 2017), Variational Autoencoders (Kingma & Welling, 2014), and Deep Ensembles (Lakshminarayanan et al., 2017).

Despite the numerous techniques to inject uncertainty estimates into machine learning, the performance of any approach is still underwhelming. Recent arguments have been made that ensembles of techniques outperform any one approach (Lakshminarayanan et al., 2017; Kuleshov et al., 2018; Guo et al., 2017; Nixon et al., 2019; Dormann, 2020). Due to the complex nature of the analytical posterior distributions, lack of complete data, prohibitive cost of training, and sensitivity to the nature of the application, an understanding of which methodology is appropriate is still in its infancy. Recently there has been some research comparing popular uncertainty quantification techniques in Deep Learning (Ashukha et al., 2020; Caldeira & Nord, 2020; Labach et al., 2019; Lakshminarayanan et al., 2017). Unfortunately, there is not much research about how these methods perform in the geosciences, where probabilities are often non-Gaussian, non-trivial, non-stationary, and high-dimensional. This paper serves as a step into answering this question by testing various probabilistic machine learning methods used for high-dimensional data with both Gaussian and non-Gaussian distributions on MLD estimation, which serves as an example problem in this respect.

Our goal for this manuscript is two-fold. First, we investigate to what extent the aseasonal variability in sea surface salinity, temperature, and height are related to, and hence useful for estimating, the aseasonal variability of the MLD. In particular, we study two geographic regions, (1) the eastern equatorial Pacific Ocean from 10S-10N and 150W-120W and (2) the southern Indian Ocean from 45S-35S from 60E-120E, over the 2011-2015 time period. As detailed in section 2, these regions are useful test cases because both are characterized by at least modest subseasonal MLD variability ($>$ 10 m subseasonal standard deviations), but the magnitudes of subseasonal variability, the climatological annual cycle, and interannual variability all differ substantially (D. B. Whitt et al., 2019). Thus, the two regions reflect useful and distinct test cases for evaluating machine learning model performance. We perform this analysis by training a series of neural network architectures to produce gridded MLD estimates using surface variables as inputs and evaluate model performance using the Argo profiles. We compare the machine learning approaches, which only use surface values as inputs, to the traditional optimal-interpolation technique that estimates using the actual MLD values from the in-situ Argo profiles. The differences in performance between the machine learning methods and optimal-interpolation schemes will reveal the extent to which the sea surface variables are useful in predicting the MLD.

Second, we focus on understanding the probability distribution of the MLD that is learned by the neural network. As a first step, we evaluate how well calibrated the neural network estimates are and what spatial and temporal patterns are revealed through sampling these distributions. We choose three probabilistic machine learning methods that cover two distinct types of uncertainty quantification: parameterization- and sampling-based methods. By evaluating these methods, we aim to understand the appropriateness of a Gaussian distribution to the data and the ability for sampling machine learning methods in exploring the posterior distribution. Finally, we compare the machine learning uncertainty quantification against uncertainty estimates from the optimal-interpolation approach. As before, this last comparison will reveal the extent to which the sea surface variables inform us about the uncertainty in the MLD.

These methods are certainly not exhaustive and so this paper is a first step to a better understanding of the aseasonal MLD variability and how machine learning can be used as a tool in this investigation. The outline of the body of the paper is as follows: first, in section 2 we detail the data and describe the data processing and methodology; second, in section 3 we describe the mathematical framework and relevant machine learning architectures that we implement; lastly, in section 4 we explain and detail the experiments and results.

2 Data

2.1 Salinity

Sea surface salinity data is the optimally-interpolated analysis of Melnichenko et al. (2016), which is an optimal interpolation of observations from the Aquarius satellite and uses corrections to minimize bias relative to in-situ data. The data exists on a $\frac{1}{2}$ degree, weekly grid spanning roughly 2011-2015 (200 weeks). A random 150 week sample constitutes the training data, with the remaining being used for testing and validation. This grid is the coarsest of all the variables and thus will form the basis that we interpolate and re-sample the other data onto. To calculate an estimate of the climatology, we calculate monthly means using only the training data, taking a 4 week boxcar moving average, binning data into months and averaging over the bins.

2.2 Temperature

Sea surface temperature data comes from the GHRSST Level 4 Global Foundation Sea Surface Temperature analysis dataset (Remote Sensing Systems, 2017). This dataset uses Optimal Interpolation (OI) from several microwave sensors. The data exists on a $\frac{1}{4}$ degree, daily grid spanning roughly 2001-2018. To calculate an estimate of the climatology, we set aside the years 2011-2015 and calculate a 4 week boxcar moving average on the remaining data. From the smoothed data, we take bins according to each month and average over the bins, resulting in an approximate monthly climatology. To calculate anomalies, we bin the 2011-2015 data into months and subtract the monthly climatology. Then, to be able to compare to the salinity dataset, we up-sample from the daily values to weekly data and optimally interpolate onto a $\frac{1}{2}$ degree grid.

2.3 Height Anomaly

Sea surface height anomaly data comes from the MEaSUREs Gridded Sea Surface Height Anomalies dataset (Zlotnicki et al., 2019). The data exists on a $\frac{1}{6}$ degree, 5-day grid spanning roughly 1992-2019. We do not calculate and remove climatologies from this data set. To be able to compare to the salinity dataset, we up-sample from the 5-day values to weekly data and optimally interpolate onto a $\frac{1}{2}$ degree grid.

2.4 Mixed Layer Depth

Argo data is available through Cabanes et al. (2013). The MLD is defined for about 1.5 million profiles of temperature and salinity that pass quality controls in the time span from 2000-2017 (D. B. Whitt et al., 2019; D. Whitt et al., 2020).

To calculate an estimate of the climatology from the individual MLD measurements, we take the years 2002-2010, and 2016-2017, bin the data into 2° latitude and 4° longitude bins, re-sample onto a daily grid and take four week moving averages in each bin. This smoothed data is then grouped into months. Both an average and standard deviation are calculated in order to compute the mean and standard deviation of the monthly climatology in each bin.[1] Anomalies are created by taking each profile from the withheld 2011-2015 Argo data and subtracting the climatology according to the profile's bin

[1] For the regions included in our studies, all bins have enough data to calculate the monthly climatology. There are many regions, such as some seas surrounding Indonesia, for instance, that do not have sufficient data.

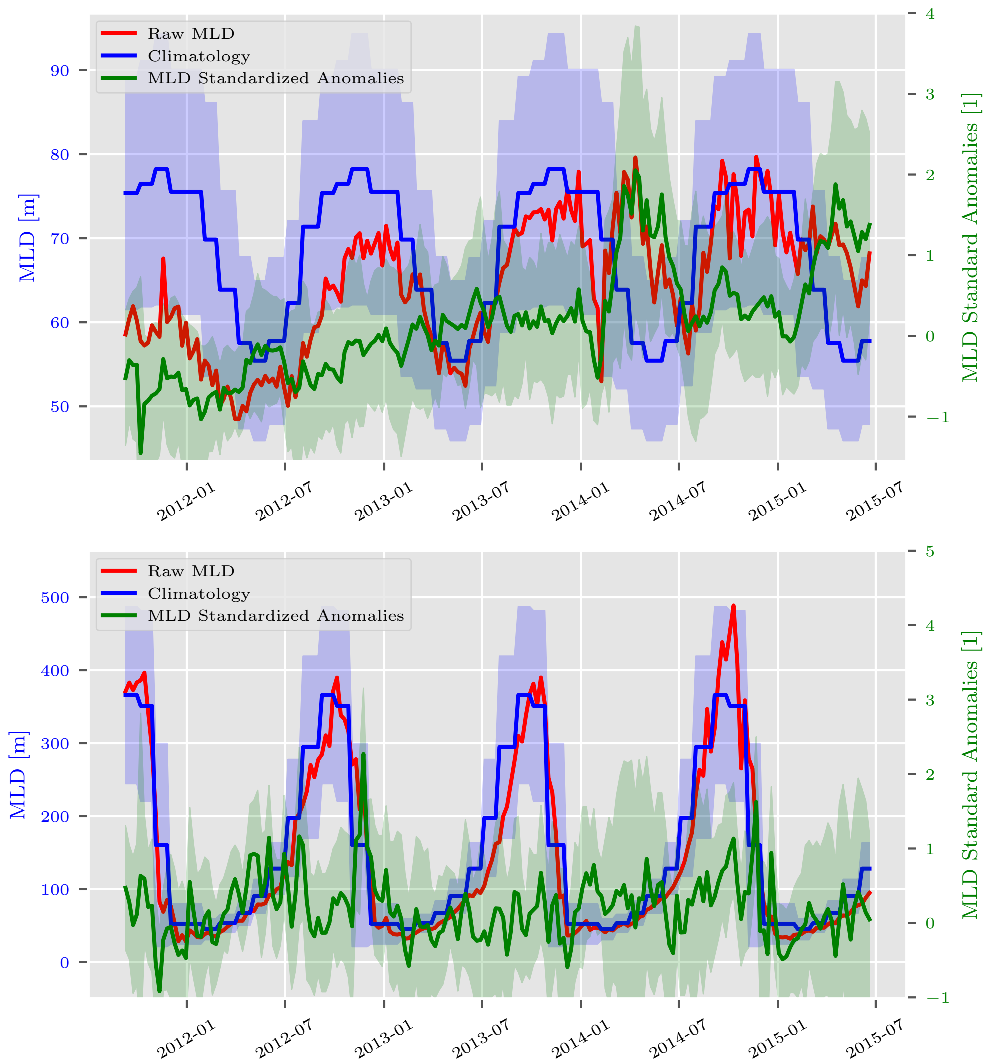


Figure 1. Several time series of the average MLD in each region at weekly resolution in the equatorial Pacific (top) and southern Indian Ocean (bottom), including the ensemble average of the MLD profiles over the domain (red), the ensemble average of the corresponding standardized MLD anomalies (green), and the area-average of the gridded monthly MLD climatology (blue). The blue shading represents the area-average of the gridded monthly standard deviations, and the green shading represents the ensemble standard deviation of the profile-wise standard anomalies. (Top) equatorial Pacific region (120W, 10S) - (150W, 10N). (Bottom) southern Indian Ocean (45S, 60E) - (35S, 120E).

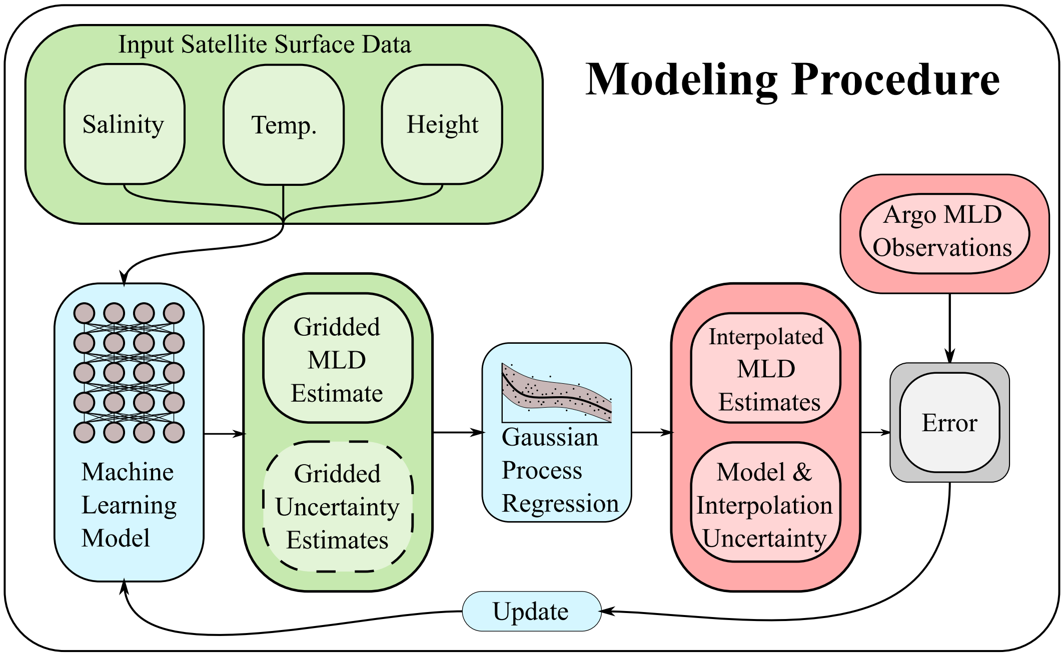


Figure 2. A schematic of the modeling procedure. Satellite sea surface data is fed into the machine learning model to produce a gridded MLD estimate (with some form of an uncertainty estimate if the machine learning model is probabilistic). To compare with the observations and optimize parameters, these gridded estimates are fed into a Gaussian process regression model (with its own hyper-parameters that are optimized) to produce MLD estimates interpolated to the locations where the Argo observations exist. These interpolated estimates are automatically associated with uncertainty estimates that derived from either just the Gaussian process interpolation uncertainty (if the model is deterministic) or a combination of the Gaussian process uncertainty with ML model uncertainty (if the ML model has uncertainty estimates). The interpolated estimates are then compared with the observations to estimate various errors.

and date. In addition, for each profile, we divide by the bin's corresponding monthly standard deviations to create standardized anomalies. Fig. 1 shows the time series of the raw MLD data, including the ensemble average of the individual profiles in each region, the ensemble average of the standardized anomalies at each profile, and the area-average of the gridded climatology, in two spatial regions under study (120W, 10S) - (150W, 10N) and (45S, 60E) - (35S, 120E). The character of the anomalies and standardized anomalies are not dissimilar, but the standardized anomalies have a more appropriate scale for machine learning purposes (see the Acknowledgements for data availability).

2.5 Evaluation Regions

In order to evaluate the behavior of the machine learning models in two different oceanic regimes, we choose to investigate two geographic regions with very different MLD variability on timescales from subseasonal to interannual but significant subseasonal MLD variability to learn in both cases. First, we choose the equatorial Pacific Ocean (10°S - 10°N and 150°W - 120°W), which has modest subseasonal MLD standard deviations ($\sim$ 15 m), a small climatological annual cycle ($\sim$ 20 m), and substantial interannual variability (see Fig. 1 and (D. B. Whitt et al., 2019)). Second, we choose to study the southern Indian Ocean (45°S - 35°S and 60°E - 120°E), which features larger subseasonal standard deviations (~ 50 m), a much larger climatological annual cycle (~ 300 m), but relatively weak interannual variability.

Hence, both regions contain substantial subseasonal MLD variability to learn, but the absolute magnitudes of the subseasonal variability as well as the relative magnitudes of subseasonal, seasonal, and inter-annual variability differ dramatically.

In order to test our framework for estimating MLD using sea surface information we perform the following experiment on each region of interest. On the 150 (out of 200 total) weeks of training data, we apply the training procedure summarized in Fig. 2 and described in more detail in section 3 (see the Acknowledgements for a link to the software).

On the remaining 50 weeks of testing and validation data the model predicts a dense grid of MLD estimates based solely on the sea surface information as input. From this dense grid, we interpolate the estimates onto the locations where in-situ Argo profile observations of the MLD exist and compute error statistics between the interpolated estimates and the observations. The interpolation is done using a Gaussian process (see section 3.1) regardless of the machine learning method. We denote this testing procedure as measuring the out-of-sample performance of the method.

3 Methods

We consider a simple but general model for the relationship between the surface variables, salinity (S), temperature (T), and height (H), and mixed layer depth model output (d),

$$d = f(S, T, H; \theta) + \sigma, \quad \sigma \sim \mathcal{N}(0, \Sigma). \tag{1}$$

where θ refers to the collection of function parameters. The surface variables exist on a pre-specified grid, $\mathbf{x}$, of total size M and the function f may generally couple surface variables from across this grid to produce d at a particular grid point. The difference between the mixed layer and the output of f, σ, is assumed to be a normally distributed random variable according to the covariance Σ that expresses the spatial uncertainties in this functional relationship. The exact structures and parameterizations of f that we use in this paper are described in section 3.2 while the methods we use to specify Σ are presented in section 3.3.

Both the functional relationship f and the covariance matrix Σ are data-driven (i.e., agnostic to the underlying physics) and informed via observations d_o that exist at arbitrary (ungridded) locations, $\mathbf{x}_o$ where freely-drifting Argo floats collect a profile. In order to couple the gridded variables with the ungridded observations, we define the relationship between our model and the observations to be a Gaussian process,

$$d_o = Ld + \nu, \quad \nu \sim \mathcal{N}(0, V), \tag{2}$$

which will be further defined in section 3.1. Importantly, L and V, the spatial projection and covariance matrices, are independent of the observation values and only depend on the observation locations, model grid locations, and model uncertainties. The Gaussian process relationship, in our study, is entirely a spatial relationship that accounts for spatial covariance between observations of the MLD. This implicitly means, however, that L and V change depending on the particular week the data is from, but only because the particular locations $\mathbf{x}_o$ where estimation and validation occurs vary from week to week.

A further consequence of the chosen relation between the observations and model (2) is that it defines the objective function, i.e. the conditional likelihood probability distribution, that will be maximized to fit the parameters of the nonlinear function f:

$$\ln p(d_o|d) = -\frac{1}{2}(d_o - Ld)^T V^{-1}(d_o - Ld) - \frac{1}{2}\ln|V| - \frac{M}{2}\ln 2\pi. \tag{3}$$

Details of this optimization procedure are given in section 3.2. Here, it is implicitly understood that d, and hence $p(d_o|d)$, is a function of the input variables S, T, H, the architecture of the function f, and the parameters of f, θ.

The Gaussian assumptions made in Eq. 1 is primarily for notational convenience. The model definition (Eq. 1) can easily be modified to include non-Gaussian noise by including a stochastic component in f, $f(S, T, H; \theta, \sigma)$. This type of noise component is important if we expect the noise to be a nonlinear function of the surface variables. To account for this possibility, two of the probabilistic machine learning methods that we test in this paper, Dropout and Variational Auto-Encoders (see section 3.3) are formally of this type and require sampling to determine the covariance for use in the Gaussian process. The Gaussian assumption made in (Eq. 2) is a reflection of the belief that the interpolating operator between the gridded locations and Argo locations is appropriately approximated by a linear function. We believe that this is not overly restrictive since most optimal interpolation techniques make similar assumptions.

3.1 Gaussian Process Regression

Gaussian Process Regression is closely related to the somewhat more general Optimal Interpolation and Kriging frameworks. For a more detailed history and exposition, see Cressie (1993). A Gaussian process is any collection of random variables for which any finite number have a joint Gaussian distribution and, as a result, is completely determined by a mean and covariance function (Rasmussen & Williams, 2006). Given a set of (2-dimensional) observation locations $\mathbf{x} = (x_1, \ldots, x_M)^T$, we define the mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ of the process $d(\mathbf{x})$ as

$$m(\mathbf{x}) = \mathrm{E}\left[d(\mathbf{x})\right] \tag{4}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{E}\left[(m(\mathbf{x}) - d(\mathbf{x}))\,(m(\mathbf{x}') - d(\mathbf{x}'))\right] \tag{5}$$

Typically the mean function is set to zero and covariance function is parameterized according to some kernel function. Various kernel functions impart different types of regularity (differentiability): the exponential kernel leads to non-differentiable outputs, the Matern Class of kernels have a regularity parameter, and the squared exponential kernel leads to smooth outputs. In our study, the squared exponential kernel,

$$k(\mathbf{x}, \mathbf{x}') = \alpha e^{-\frac{1}{2\ell}\|\mathbf{x}-\mathbf{x}'\|^2} + \beta \tag{6}$$

where α and ℓ are hyperparameters that control the amplitude and length-scale of the corresponding covariance structure, was chosen because of its marginally better performance and efficiency compared to Matern class kernels. We train our Gaussian process hyperparameters by optimizing according to the Gaussian process prior probability distribution over the training observation points $\mathbf{x}$,

$$\ln p(\alpha, \ell, \beta|d) = -\frac{1}{2}d^T K(\mathbf{x}, \mathbf{x})^{-1} d - \frac{1}{2}\ln|K(\mathbf{x}, \mathbf{x})| - \frac{M}{2}\ln 2\pi, \tag{7}$$

where the covariance matrix has entries $K_{i,j}(\mathbf{x}, \mathbf{x}) = k(x_i, x_j)$. To regularize the optimization process and ensure positivity of α, ℓ, and β, priors are occasionally placed on the hyperparameters in a Bayesian fashion. In our study, this type of implementation had minimal impact on the optimized values. In circumstances where either computational considerations are not a concern or available training data is limited, it is also possible to optimize the hyperparameters by cross-validating and minimizing the conditional likelihood distribution, for details see Rasmussen and Williams (2006). The variance hyperparameter β can, in general, be made anisotropic at the expense of increasing the total number of hyperparameters, but we do not consider such options in this study.

During the training of the neural network, i.e. while optimizing the parameters in f via Eq. 3 using backpropagation on training data from a given week, the Gaussian process hyperparameters must be re-optimized according to Eq. 7 because the Gaussian process parameterization depends on the Argo profile locations (and model covariance Σ, if available) which generally vary from one training week to the next.

Once the Gaussian process has been optimized using function values $(\mathbf{x}, d)$, we can perform inference at the Argo spatial locations $\mathbf{x}_o$ to obtain estimates of d_o. The inference procedure follows Eq. 2 with L and V given by the equations

$$L = k(\mathbf{x}_o, \mathbf{x})\left(k(\mathbf{x}, \mathbf{x}) + \Sigma\right)^{-1} \tag{8}$$

$$V = k(\mathbf{x}_o, \mathbf{x}_o) - k(\mathbf{x}_o, \mathbf{x})(k(\mathbf{x}, \mathbf{x}) + \Sigma)^{-1} k(\mathbf{x}, \mathbf{x}_o). \tag{9}$$

Thus, the trained kernel function is independent of time and depends only on distance $\|\mathbf{x}-\mathbf{x}'\|$ not location $\mathbf{x}$ or time, but L and V depend on location and time because Σ depends on location $\mathbf{x}$ and the particular points chosen for estimation $\mathbf{x_o}$ (e.g., the Argo profiles locations) vary with time.

3.2 Machine Learning

The main objective of this paper is to learn a relationship between the sea surface variables (salinity, temperature, height) and mixed layer depth. Without an a priori physics-based model, one must choose a reasonably parameterized model to approximate this relationship. Traditionally this relationship is represented via some linear or simple non-linear parameterization where one hopes that the true relationship lies in, or is not too far from, the output space of the model. For example, a basic linear model that we test in this paper is of the form,

$$d_\ell = \begin{bmatrix} c_1(\mathbf{x}) \\ c_2(\mathbf{x}) \\ c_3(\mathbf{x}) \end{bmatrix} \cdot \begin{bmatrix} S \\ T \\ H \end{bmatrix} + b + \sigma, \quad \sigma \sim N(0, \Sigma) \tag{10}$$

Such models, however, are typically not expressive enough to represent arbitrary relationships. The revolution of machine learning, and, in particular, deep learning, has been born out of the need to express arbitrary functional relationships amid a dearth of observational data. While there exists several popular machine learning architectures, we base our paper around modifications of the quintessential deep learning model, the feedforward neural network (FNN) (Goodfellow et al., 2016). FNNs are represented by composing together many different functions in series to form a chain,

$$f(x) = f^{(n)}(f^{(n-1)}(\cdots f^{(1)}(x)\cdots)), \tag{11}$$

$$f^{(i)}(x) = a\left(x^T W_i + b_i\right), \tag{12}$$

where W_i is a matrix of weights, b_i is a bias term, and $a(\cdot)$ is what is referred to as an 'activation function', that applies a simple non-linearity element-wise to the affine transformation of the input, x. Common examples of activation functions include the sigmoid, softplus, and rectified linear functions. Based on the experiments in Gal (2016), we implement the rectified linear unit as the activation function in all of our neural network layers, although it is possible that, among all of the available activation functions, another function would result in superior performance. We will denote the collection of neural network parameters as $\theta = \{W_1, \ldots, W_n, b_1, \ldots, b_n\}$.

The training of a neural network entails obtaining an estimate of the parameters, $\hat{\theta}$, by approximately solving the optimization problem,

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \ln p(d_o|d) \\ &= \arg\min_\theta \left\{ g(\theta) - \sum_{j=1}^{n_{\text{train}}} \ln p_j(d_o|d) \right\} \end{aligned} \quad (13)$$

where $g(\theta)$ is a regularization function that is applied to both constrain the possible parameter values and stabilize the optimization procedure. As written, $p_j(d_o|d)$ refers to the joint probability distribution between the j^{th} input and output data. The optimization procedure includes all training data but, in practice, subsetting is common (as in batch gradient descent (Ruder, 2016)). We only seek an approximate solution to Eq. 13 for two reasons: first, the optimization problem is highly-non trivial, non-convex, and high-dimensional with many local minima and obtaining a global minimum is infeasible; second, the ultimate goal is for the parameters to lead to a function f that generalizes well to data not in the training set and over-training might ultimately hinder this goal (Caruana et al., 2001). The problem of over-fitting and poor generalization is one of the largest obstacles to good machine learning performance, particularly in applications where prediction involves extrapolation beyond whatever data was in the training set. All of the neural networks implemented for this paper are done using the TensorFlow and TensorFlow Probability frameworks (Abadi et al., 2016; Dillon et al., 2017).

Because our study is limited to only 150 training weeks, we implement a non-standard training strategy to help reduce overfitting. For each epoch (a single run through the entire training data) we divide the 150 training weeks randomly into 6 batches of 25 weeks. The first batch is held out and the current loss on that batch is saved. For each subsequent batch, the loss for that batch is used to update the model parameters. To update the parameters, we use the Adam optimizer with initial learning parameter set to $1e-3$ (Kingma & Ba, 2015). With the updated model parameters, we calculate a new loss on the first, held-out batch. If that new loss is less than the saved loss, then the updated parameters are accepted and the new loss is saved. If the new loss is larger than the saved loss then the parameters are only accepted with

$$\text{probability of acceptance} = \exp\left(\text{saved loss} - \text{final loss}\right).$$

This training strategy reduces the amount of overfitting because it forces updates to be generalizable to the held out batch, which acts as a 'testing batch'.

FNNs with enough hidden layers have been proven to serve as a universal approximator (Hornik et al., 1989; Cybenko, 1989; Leshno et al., 1993). This means that, at least theoretically, there exists a FNN that can represent whatever functional relationship exists between the sea surface variables and MLD. Unfortunately, there is no guaranteed way to find this optimal relationship. While the optimization problem (Eq. 13) has a natural inherited probabilistic framework, even an exact solution has no guarantee of agreeing with the 'true' relationship. The construction of these optimization frameworks and the regularization functions is often done by trial and error since there is, as of yet, no clear casual relationship between tuning the architecture settings and the resulting uncertainty estimate - even if the model can be viewed through a (Bayesian) probabilistic framework.

Finally, since the (approximate) solution to Eq. 13 is not accompanied with natural uncertainty estimates for the parameters, it can be difficult to obtain calibrated probabilistic estimates of $\hat{d}$. To truly obtain samples from the posterior $p(d|d_o, S, T, H, \theta)$, we would need to incorporate any and all uncertainties that exist in the inputs, observations, model parameters, and model framework and be able to sample from them effectively. Due to the high-dimensionality of the problem, this is computationally infeasible and therefore we must rely on adequate approximations. In the next section, we outline the approximations that we test in this manuscript.

3.3 Probabilistic Machine Learning Models

The simplest technique to introduce uncertainty estimates into a neural network is to implement Dropout (Hinton et al., 2012; Srivastava et al., 2014). Acting as a layer of the network, Dropout randomly sets inputs to zero at a particular rate and scales the rest of the inputs by $1/(1 - \text{rate})$. Mathematically,

$$f^{(i)}(x) = \frac{1}{1-p} M \odot a(x^T W_i + b_i), \quad M_j \sim \text{Bernoulli}(p), \tag{14}$$

where $\odot$ means element-wise multiplication. Each run of the model then has a different combination of weights that are set to zero. While originally this technique was used to reduce overfitting, it can also be viewed through a Bayesian probabilistic lens (Maeda, 2014). Running the model multiple times creates an ensemble that can be used to calculate moments of the output distribution, and, in particular, Σ and μ. It has been shown that the expected distribution from a neural network utilizing Dropout forms a Gaussian mixture distribution (Gal & Ghahramani, 2016). Therefore, there is some reason to believe that the regularity of the data distribution dictates how useful Dropout can be in uncertainty quantification.

The next simplest probabilistic technique, what we call the Variational Artificial Neural Network (VANN), also known as a heteroscedastic network, is to parameterize the output of the neural network according to some distribution. For a Gaussian distribution, for example, the output of f is a stacked vector of the mean and covariance estimates,

$$f(S, T, H; \theta) = [\mu; \text{vec}(\Sigma)], \tag{15}$$

where $\text{vec}(\Sigma)$ is the flattened covariance matrix, such that $d \sim N(\mu, \Sigma)$. This technique is relatively easy to implement with care needed to ensure that constraints on the parameters are enforced. Typically, a Bayesian framework would then impose prior probability distributions onto μ and Σ. In particular, in addition to the Gaussian likelihood, it is common to impose a Gamma or LKJ - uniform over the space of covariance matrices - prior on the covariance to prevent unnecessary shrinkage. In a feedforward neural network, this parameterization increases the number of outputs and hence the overall total number of parameters. If the number of grid points of $d(\mathbf{x})$ is M then a full covariance matrix would require $M(M+1)/2$ parameters and the corresponding number of parameters required in the neural network makes it computationally prohibitive as k grows large. To limit the computational cost, we make a diagonal assumption about the covariance to reduce the number of parameters at the expense of losing covariance information between MLD values at different grid points. Parameterization of the data distribution is not always possible if a good approximation or transformation to an appropriate probability distribution is not known and the effectiveness of this technique is reflection of the quality of that assumption.

The final method that we test is the variational auto-encoder (VAE) (Kingma & Welling, 2014). A typical VAE consists of two dense networks: an encoder that projects the inputs into a lower-dimensional latent space, parameterized by a probability distribution, and a decoder that inverts this projection and produces the original input. The loss between the decoder's output and the original system drives the learning process. A VAE supposes a prior distribution over the latent variable z, $p(z)$, that, along with the decoder network that induces a conditional likelihood distribution $p(S, T, H|z; \theta)$, forms a posterior distribution,

$$p(z|S, T, H; \theta) \propto p(z)p(S, T, H|z; \theta)$$

This posterior distribution is typically intractable and thus replaced by a variational approximation $q(z|S, T, H; \theta)$. This approximation includes a parameterization of the prior and likelihood distributions, typically Gaussian distributions with parameters that are

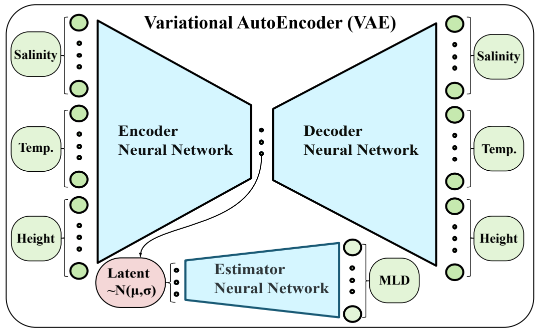


Figure 3. A schematic of the modified VAE. Training is informed by the decoder and estimator networks losses. For a full description of the training procedure for a typical VAE, see Kingma and Welling (2014).

learned in the encoder network. In our design we also use a Gaussian distribution in the latent space, and, as demonstrated in Figure 3, we couple this network with a third dense network, which we call the estimator, that transforms the latent space into an estimate of the MLD associated with the surface salinity, temperature, and sea height anomaly encoder inputs.

While the prior and likelihood distributions in a VAE are specified as Gaussian, the distribution of the output of the estimator network, that is, the MLD outputs, is not parameterized. While the difference between the MLD estimates and the MLD observations is modelled as a Gaussian process regardless of neural network architecture, the possible benefit of our chosen VAE approach is that it can produce theoretically arbitrary probability distribution $p(d|S,T,H;\theta)$. Another theoretical benefit to this approach is that, since the neural network can learn an efficient lower-dimensional representation of the inputs that capture dominant patterns, the estimator might be better able to generalize and less sensitive to small perturbations and noise in the inputs.

We summarize the ways in which the MLD uncertainty, represented as Σ, is estimated. For the non probabilistic methods (linear model, artificial neural network), there is no associated Σ. For the Variational Artificial Neural Network (VANN), Σ is a direct output of the neural network and the weights that produce this Σ are trained as in Eq. 13. For the Dropout network, each output of the network is a draw from a random distribution. Σ is the sample covariance matrix of 100 random samples from this distribution. Similarly, for the variational auto-encoder (VAE), Σ is the sample covariance matrix from 100 random outputs of the VAE network.

4 Experimental Results

We test 6 different methods on each experiment, five of which we consider as part of the machine learning framework: the linear model (Eq. 10), the feedforward artificial neural network (Eq. 11), feedforward neural network with parameterized distributional output (Eq. 15) feedforward neural network with Dropout (Eq. 14), and a variational auto-encoder. We collectively shorthand these to be 'Linear', 'ANN', 'VANN', 'Dropout', and 'VAE'. While the models presented in this study are based on the basic feedforward neural network architecture, we also tested (with poor performance) convolutional neural networks with a multitude of architectures and hyperparameters. Finally, in order to compare these methods to a traditional interpolation only approach, we implement an Ordinary Kriging scheme, which we call 'OI' for optimal interpolation, with a (spatial) spherical kernel chosen via cross-validation and parameters optimized via maximum likelihood. The OI approach only uses the in-situ MLD standard anomaly observations, with no sea surface information, to make gridded estimates. Therefore, even during the out-of-sample prediction experiments, the OI's error statistics for a given week are calculated using only that week's data. In particular, we use a cross-validation approach using a 75-25% train-test split to estimate these error statistics.

We use 3 metrics in our testing: root mean squared error (RMSE), Pearson correlation coefficient, and probabilistic calibration. These metrics are applied to modeled standardized MLD anomalies at the validating Argo profile locations (see section 2 for details). We use the typical definition of root mean squared error,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|(d_o)_i - L(d)_i\right|^2}. \tag{16}$$

RMSE is a convenient metric in that it captures the mean prediction error, but it doesn't necessarily tell us much about the relationship between the predictions and observations and it also fails to capture meaningful information about the uncertainty of the predictions. To compensate for the first deficiency, we rely on the Pearson correlation coefficient (correlation) to provide insight into the existence of (linear) relationships between predictions and the Argo MLD data. For reference, correlation is defined as

$$\text{Correlation} = \frac{\sum_{i=1}^{n}\left(L(d)_i - \overline{L(d)}\right)\left((d_o)_i - \overline{d_o}\right)}{\sqrt{\sum_{i=1}^{n}\left(L(d)_i - \overline{L(d)}\right)}\sqrt{\sum_{i=1}^{n}\left((d_o)_i - \overline{d_o}\right)}} \tag{17}$$

Common metrics that capture probabilistic calibration include skill scores such as the Brier score or the Kolmogorov–Smirnov statistic. Here, for simplicity, convenience, and data-limitation reasons, we use the following measure for probabilistic calibration,

$$\text{Calibration} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left[\left|(d_o)_i - L(d)_i\right| < \sqrt{V_{ii}}\right], \tag{18}$$

where V_{ii} is the ith diagonal entry of the covariance matrix of the Gaussian process regressor (Eq. 8) and $\mathbb{1}$ is 1 if the argument is true and 0 otherwise. Calibration is then a number between 0 and 1. It is important to remember that V also includes the covariance estimate from the probabilistic machine learning models, Σ. For non-probabilistic machine learning model, V does not include any model uncertainty beyond the learned hyperparameter β in Eq. 6. For a Gaussian statistic, the Calibration is theoretically $\approx$ 0.68, the optimal score for this metric. If a model scores lower than that theoretical threshold, it is underestimating the amount of uncertainty in the data. Conversely, a higher Calibration than the theoretical threshold represents an overestimation of the uncertainty.

We aim to give an overview of the main results from our studies. We focus on the aforementioned metrics as we compare model performance overall, and broken down by groups representing different levels of standard deviation in the observations. These metrics indicate 3 conclusions: 1) Model performance is superior in the equatorial Pacific Ocean than the southern Indian Ocean, 2) the probabilistic machine learning methods outperform traditional OI, particularly in terms of correlation and calibration, and, therefore, 3) the relative performance of machine learning algorithms indicate that surface variables can provide meaningful information about the mixed layer depth and produce estimates that are as good or better than OI methods that directly use MLD data. Finally, we visually compare the model outputs in two case studies that represent the best and worst model performance. To provide context and further applications, we also include

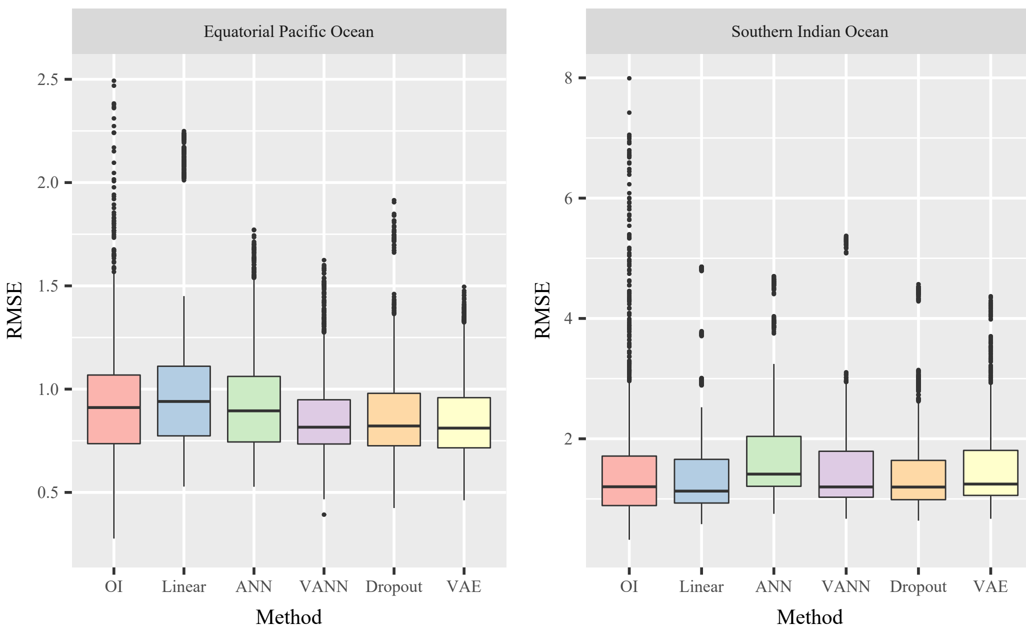


Figure 4. Root Mean Squared Errors (RMSE) on temporal out-of-sample prediction (in meters). Errors are calculated on 50 withheld validation weeks. Boxes capture 25-75% of the weekly errors with the middle line representing the median error. Dots are considered outliers - values which are 1.5× lower/upper quantile. (Left) The equatorial Pacific region (120W, 10S) - (150W, 10N). (Right) The southern Indian Ocean region (45S, 60E) - (35S, 120E). Note the difference in scales between the two regions. OI errors are calculated using cross-validation within each week (see text for details).

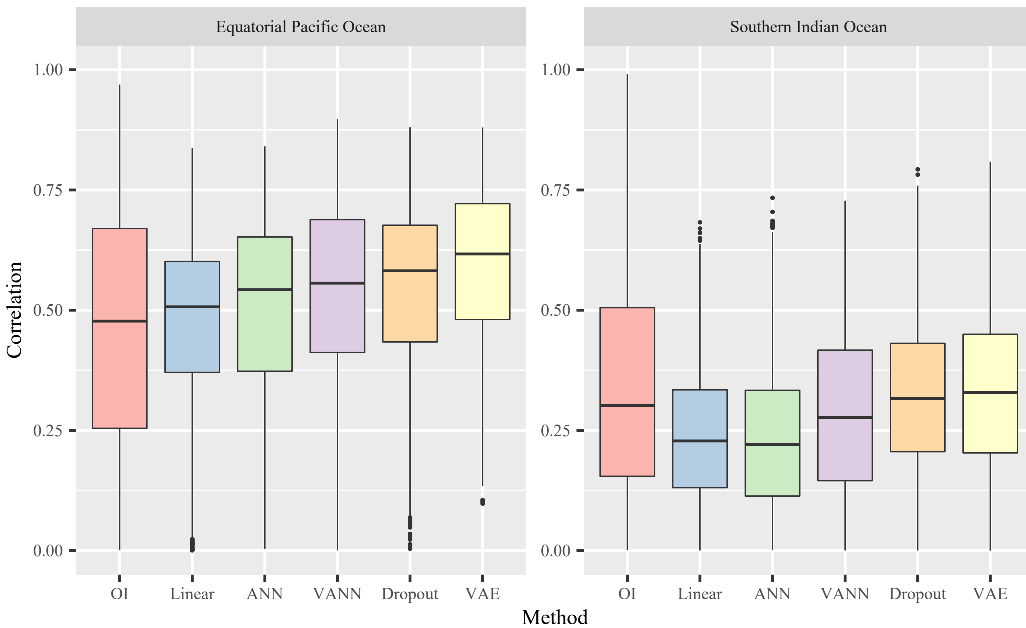


Figure 5. Correlation on temporal out-of-sample prediction (in meters) as in Fig. 6. Correlations are calculated on 50 withheld validation weeks. Boxes capture 25-75% of the weekly correlation with the middle line representing the median correlation. Dots are considered outliers - values which are 1.5× lower/upper quantile. (Left) The equatorial Pacific region (120W, 10S) - (150W, 10N). (Right) The southern Indian Ocean region (45S, 60E) - (35S, 120E). OI values are calculated using cross-validation within each week (see text for details).

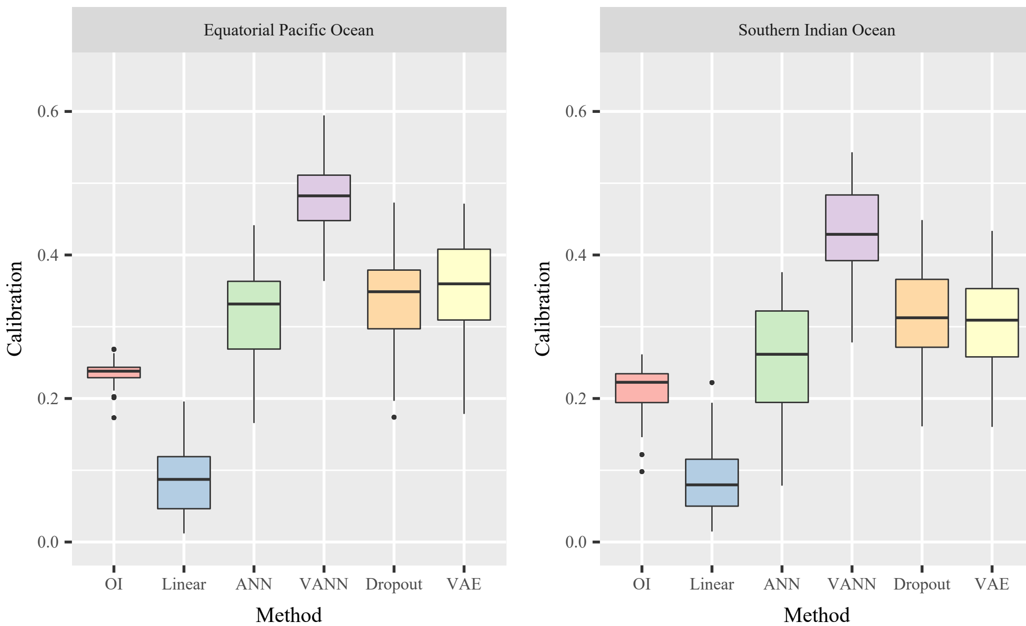


Figure 6. Measure of probabilistic calibration on temporal out-of-sample prediction as in Fig. 6. Calibrations are calculated on 50 withheld validation weeks. For each week, we find the percent observations that fall within 1 standard deviation of forecast ensembles. For a Gaussian distribution, this probability should be approximately 0.68, with greater relative values representing under-confident and lesser relative values representing overconfident predictions. OI calibrations are calculated using cross-validation within each week. (Left) The equatorial Pacific region (120W, 10S) - (150W, 10N). (Right) The southern Indian Ocean region (45S, 60E) - (35S, 120E).

model outputs from the HYCOM + NCODA Global 1/12° Analysis (Fox et al., 2002; Cummings, 2006; Cummings & Smedstad, 2013) for a visual comparison with our purely data-driven approaches.

Considering first the RMSE, model performance is superior in the equatorial Pacific Ocean compared to the southern Indian Ocean, and the various models differ only modestly within each region. Fig. 4 (note the difference in scales of the vertical axis) shows the RMSE results over the two regions. The machine learning methods seemingly perform well against OI, particularly in the equatorial Pacific as the Dropout and VAE methods have the lowest median RMSE and 25% - 75% range. In the southern Indian Ocean, the Linear method performs well, initially suggesting that the mean dynamics can be well approximated by a linear combination of the surface variables. The number and range of OI outliers, in comparison to machine learning approaches, demonstrates that the machine learning approaches offer more stable predictions.

The correlation analysis underscores and further confirms the result (derived from RMSE above) that the overall model performance is better in the eastern equatorial Pacific Ocean compared to the southern Indian Ocean (Fig. 5). However, the correlations also reveal more substantial differences between the models in each region. In the equatorial Pacific, it is clear that the machine learning methods perform better than traditional OI, with the VAE performing the best. In the southern Indian Ocean, however, there is little separating the performance between OI and probabilistic machine learning methods, although the VAE is marginally the best performing model in this region as well. A key difference between the RMSE results in Fig. 4 and the correlations in Fig. 5 is that the linear method, while having a small predictive RMSE, has poor correlation with the observations. From other testing, we believe that the linear model has both small RMSE and correlation because the outputs of the linear method are generally smaller values.

The calibration results in Fig. 6 demonstrate that the probabilistic machine learning approaches using surface data are significantly better at estimating the posterior uncertainty than OI and MLD data alone. Furthermore, model performance is again superior (albeit modestly so) in equatorial Pacific Ocean compared to the southern Indian Ocean. The linear model performs very poorly in comparison to the other machine learning methods. The traditional OI approach also has poorer performance compared to the machine learning models. In addition, all probabilistic techniques appear to perform slightly better than the non-probabilistic ANN (in terms of both calibration and RMSE). However, the smallness of the differences between ANN and the other ML models suggests that much of the uncertainty manifest in all the ML model calibrations is due to the Gaussian Process regression, since the ANN does not have inherent MLD uncertainty estimates. Among the three probabilistic machine learning models, VANN, Dropout, and VAE, the VANN has dramatically better calibration than the other two methods. This discrepancy shows that, in these particular case studies, explicitly parameterizing the noise better captures the underlying uncertainty than the sampling-based approaches.

The conclusion from the calibration metric are mirrored in Fig. 7, where the empirical cumulative distribution of the models is plotted against the distribution of the observations. The diagram represents the Lines closer to the optimal red line in that figure represents better model calibration. It is clear from this plot that the VANN and VAE have superior performance in estimating the tails of the distribution when compared to other methods and the OI. It is true, however, that overall performance is lacking. The behavior of each line indicates that the tails of the model distribution are shorter than the observational distribution - another indication that extreme MLD values remain difficult for the models to predict.

The difference between performance in VANN vs. Dropout and VAE could plausibly explained by suggesting that the posterior probability distribution of the MLD given

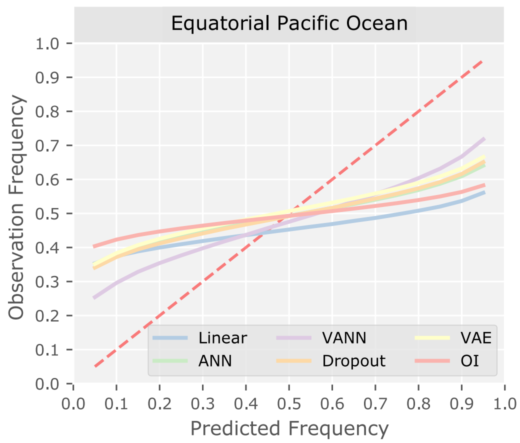
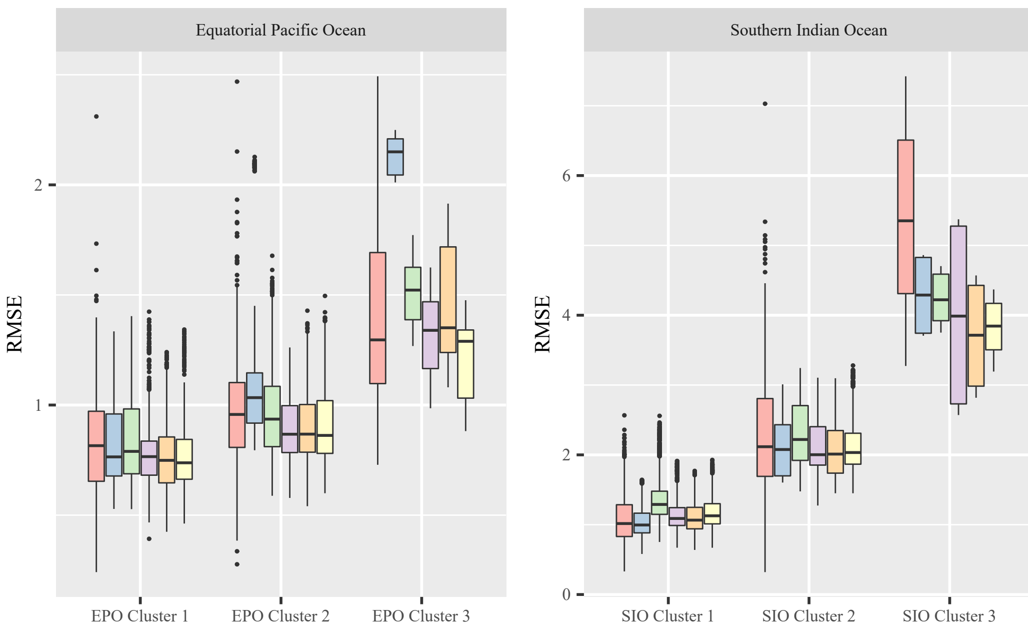


Figure 7. Probability plot comparing the empirical cumulative distributions of the model outputs against the data. The dotted red line would represent perfect agreement between models and observations. A value above and to the left of the red line indicates a part of the distribution that is over-represented, whereas a value below and to the right of the red line indicates a part of the distribution that is underrepresented.

satellite data is closely approximates a Gaussian distribution and hence well estimated by the VANN. Alternatively, the available data may not be sufficient to allow the sampling-based methods (Dropout and VAE) to learn the posterior distribution.

To reveal how the model performance depends on the MLD variability, we group the observed MLDs at the Argo profile locations by the (ensemble) standard deviation of all observed standardized MLD anomalies (defined in section 2.4) in the same week and region using K-Means clustering. We find that model performance generally degrades in terms of RMSE (Fig. 8) but improves in terms of correlation (Fig. 9) in weeks with higher standard deviations. But, model calibration (not shown) is relatively insensitive to the weekly variability of MLD anomalies. With regard to RMSEs in Fig. 8, we find that the increases in RMSE with standard deviation are fairly consistent across the models, and the slope RMSE-over-standard-deviation is roughly 1 in both regions. In addition, the probabilistic machine learning models have about equal or smaller RMSE than the OI at all levels of variance. Finally, it is notable that for the weeks with the largest observation standard deviations, the OI has particularly large RMSEs in the southern Indian Ocean, whereas the linear method has particularly large RMSEs in the equatorial Pacific.

With regard to the correlations in Fig. 9, we find that the increasing standard deviation of the observations in the equatorial Pacific Ocean improves model performance to a much greater degree than in the southern Indian Ocean. Interestingly, the comparisons between the models within each standard deviation cluster qualitatively mirror those of the whole dataset (c.f., Figs. 9 and 5): machine learning models generally produce higher correlation than OI, particularly in the equatorial Pacific Ocean. The only notable exception is the bin with high standard deviation in the Southern Indian Ocean, where the VANN, Dropout and VAE models have notably higher correlation than the other methods, while OI performs particularly poorly. Finally, the relatively high correlations at large standard deviation in the equatorial Pacific suggest, potentially, that the dynamics that cause large mixed layer depth anomalies also strongly couple with the surface variables in this region.



Cluster	# of Weeks	Min SD	Max SD
EPO 1	24	0.38	0.66
EPO 2	23	0.68	1.02
EPO 3	3	1.22	1.51

Cluster	# of Weeks	Min SD	Max SD
SIO 1	36	0.49	1.46
SIO 2	12	1.64	2.77
SIO 3	2	3.66	4.65

Figure 8. RMSEs divided by region and clustered by the standard deviations of the ensembles of MLD standard anomalies in a given week in (Left) the equatorial Pacific Ocean and (Right) the southern Indian Ocean. (Bottom) Table showing the number of weeks in each cluster, the minimum standard deviation in each cluster, and the maximum standard deviation in each cluster. (Top) The distribution of the RMSE for each method, corresponding to 40 samples from the posterior distribution for each week, separated by cluster. The boxplots are colored as in Fig. 4. Note the difference in scales between the two regions.

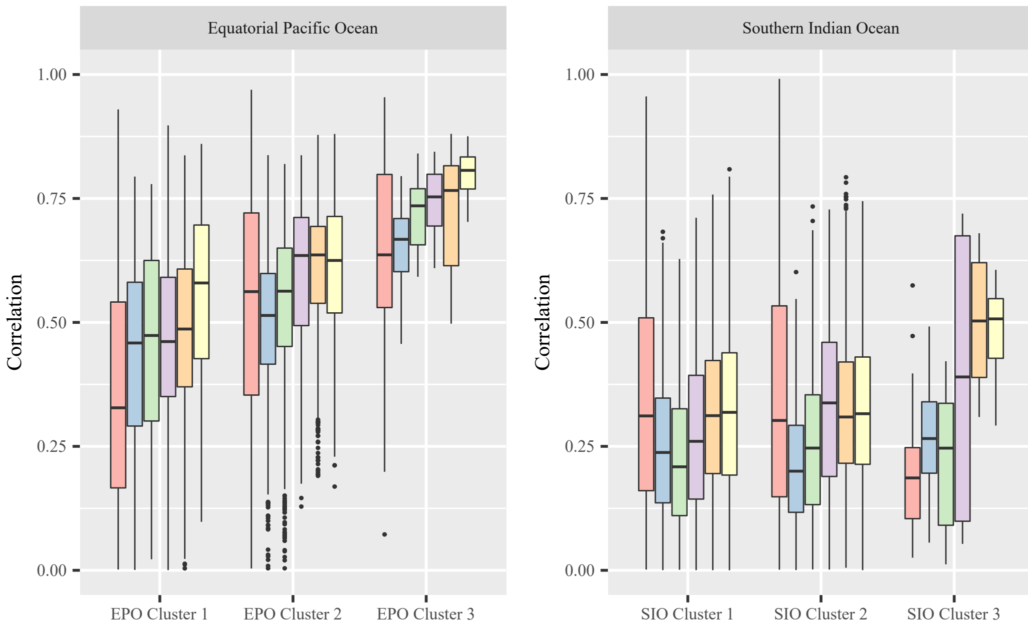


Figure 9. As in Fig. 8, but correlations instead of RMSE.

Taken together, the results indicate that, in the equatorial Pacific Ocean and to a lesser extent in the southern Indian Ocean, the surface information provides just as, if not more so, valuable information in estimating the MLD as the existing Argo observations of the MLD.

To give a visual and spatial sense of the range of model estimates, we demonstrate two extreme ends of the prediction spectrum, the worst and best predictive weeks for our models. Full model output for all available weeks are available online (Foster et al., 2020) at `https://www.doi.org/10.5281/zenodo.4421752`. The week corresponding to the worst RMSE performance is the week of 11-23-2012 in the southern Indian Ocean. If you compare this with Fig. 1, this corresponds to a period of particularly large anomalies. The average RMSEs for this particular week corresponding to the OI and VAE models are approximately 6.14 and 4.16. Similarly, the week of the relative best performance (now in terms of correlation coefficient) is 05-09-2014 in the equatorial Pacific Ocean. The corresponding average correlations (and RMSEs) for the OI and VAE methods are 0.68 (1.09) and 0.83 (0.99). Figs. 10 and 11 show A. the data with overlaid sea level height contours, B. smooth gridded climatology, C. standard anomaly OI model output, D. VAE model output, E. reanalysis of VAE model output and observations, and F. HYCOM+NCODA Global 1/12° Analysis for these two weeks. MLD values are derived from the HYCOM+NCODA reanalysis by applying the MLD definition in D. B. Whitt et al. (2019) and averaging over the appropriate week (the raw southern ocean HYCOM data is 3-hourly and the equatorial Pacific data is daily). Each of the machine learning and OI model outputs are computed as MLD standard anomalies and are transformed back to MLD estimates for plotting. Because the output of the VAE model does not use observations at prediction time, we can perform our own reanalysis by finding the minimum of the associated posterior distribution,

$$\begin{aligned}\hat{d} &= \arg\min_d \quad -\ln p(d|d_o, d_{\mathrm{m}}),\\ &= \arg\min_d \; (d-d_{\mathrm{m}})^T\Sigma^{-1}(d-d_{\mathrm{m}}) + (Ld-d_o)^T V^{-1}(Ld-d_o).\end{aligned} \tag{19}$$

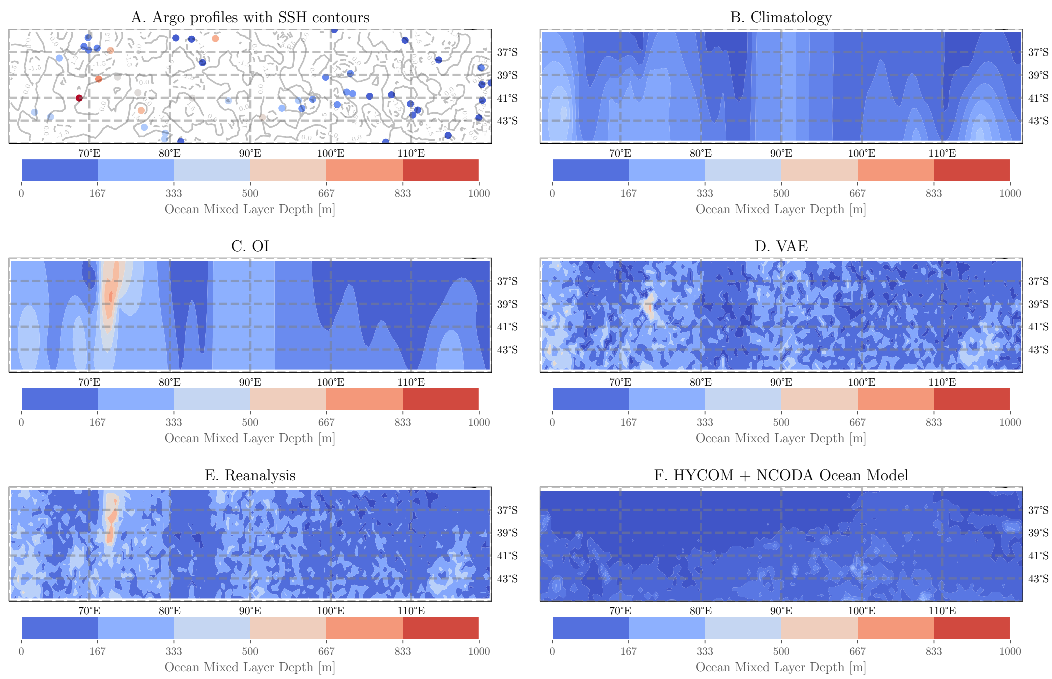


Figure 10. MLD estimates, estimated on standard anomalies with climatologies added back in, corresponding to the date of worst RMSE, achieved by VAE approach in the southern Indian Ocean, 11-23-2012. Methods from top left to bottom right: A. Argo float observations with sea level height contours of 0.5 meters are overlaid (blue is lower height), B. smooth gridded climatology, C. optimally interpolated standard anomalies with climatologies, D. VAE model with climatologies, E. Reanalysis of VAE and observations, and F. HYCOM+NCODA ocean model - see text for more details.

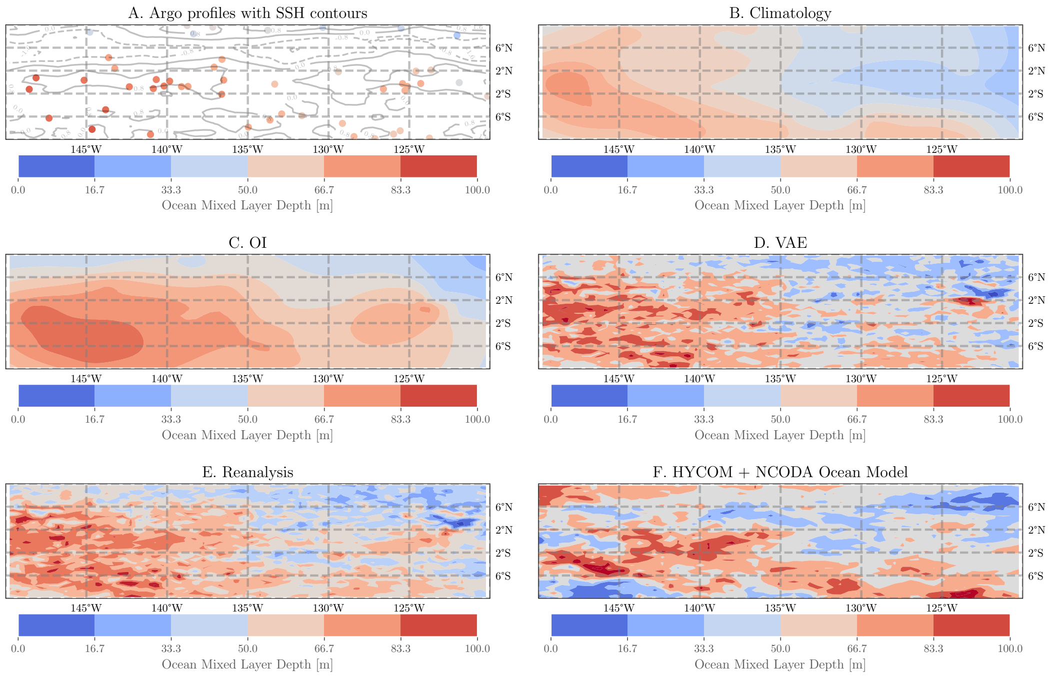


Figure 11. MLD estimates, estimated on standard anomalies with climatologies added back in, corresponding to the date of best average correlation, achieved by VAE approach in the equatorial Pacific Ocean, 05-9-2014. Methods from top left to bottom right as in Fig. 10

In Fig. 10, the week representing the collectively worst model performance, is an example of an extremely large MLD standard anomalies that can occur in late spring due to a delay in the springtime transition from deep winter to shallow summertime MLDs, as seen in Fig. 1. In this week, there is a narrow cluster of abnormally large MLD values that are visible in panels A, C, D and E. The OI model outputs are visually smooth, as a result of the spherical kernel used to do the interpolation, but underestimate the magnitude of the data. The VAE model output, as a result of being a function of the sea surface data, contains many small scale features that create a visually noisy gridded estimate. In addition, there are clusters of large anomalies where the data does not suggest any (near 115°E and 43°S for example). The reanalysis, as a result of being a variance-weighted average between the VAE and the observations, more closely resembles the OI estimate but still contains much more small scale variability. In the HYCOM + NCODA Global reanalysis, the model does not seem to capture the large MLD values that are seen in the Argo data, which might be due to the relative uncertainties in the HYCOM + NCODA Data Assimilation procedure. Direct comparisons between the VAE reanalysis and HYCOM+NCODA model should not be over-exaggerated because the differences in variance specification.

Similar to the worst case, the best case (achieved by the VAE model) occurs in a week of large standard anomalies in the equatorial Pacific (Fig. 1). As opposed to the worst case study, in this case study (Fig. 11) the climatology offers a lot of structure that is manifested in the MLD that week. The OI model output presents a very spatially coherent MLD estimate. The machine learning models, as a result of being functions of the sea surface inputs, have smaller scale features that modify the overall structure of the gridded MLD. The VAE model output, while having better performance in estimating the MLD standard anomalies than the OI at the observation locations, appears to have a greater stratified estimate. That is, the VAE model seems to overestimate the magnitude of standard anomalies. The reanalysis of the VAE model output and observations retains a mixture of the smaller scale feature from the VAE model and the coherent structure apparent in the OI output. The HYCOM + NCODA reanalysis closely captures the scale of the Argo MLD values, but the overall structure does not visually seem to match the observations. Again, the comparison with the HYCOM + NCODA reanalysis should be taken with appropriate qualification.

5 Conclusion and Discussion

The ocean mixed layer interacts with the atmosphere and deep ocean on a multitude of spatial and temporal scales. Heat exchange between these bodies has significant impact on subseasonal and interannual (aseasonal) timescales and can influence the behavior of dominant modes of variability (i.e. ENSO, MJO, tropical cyclones). Proliferation of Argo floats have dramatically increased the number of observations of the ocean over the preceding decades but are still too sparse to resolve fine spatio-temporal features of the MLD. Satellite data, however, is able to provide fine resolution gridded maps of sea surface variables, but cannot provide subsurface information.

The first goal of this work was to analyze the extent to which satellite data of sea surface variables can provide information useful for estimating the MLD. We built several machine learning models to learn such a relationship based on available data. We found that in terms of both root mean squared error, correlation, and probabilistic calibration, the machine learning model results suggest that the satellite data is equally if not more useful in estimating MLD values and uncertainties than MLD observations alone, given that sufficient MLD observations are available for out of sample training (Figs. 4 & 6). The exact relative performance between these methods can depend on the location of interest and the aseasonal variance, but we believe that the machine learning methodology can be widely applicable and competitive with optimal interpolation approaches globally. In particular, the Argo mixed layer depth samples with increased variance in the equatorial Pacific Ocean, whose subannual variability includes a relatively strong aseasonal component, seem to be more strongly connected with the surface dynamics. Therefore, including surface information together with in-situ MLD estimates may be useful for generating improved reanalyses of the upper ocean under these circumstances. The second goal of this work was to use sophisticated probabilistic learning approaches to better understand the probability distribution of the MLD. The probabilistic approaches capture uncertainty to a greater extent than the optimal interpolation approach, but it is clear that, whether because of data or model limitations, more work is needed to obtain truly calibrated posterior probabilities. While initial results suggest that a Gaussian approximation of the conditional posterior distribution is appropriate, insufficient data might also explain the relative under-performance of the sampling-based probabilistic machine learning methods that we tested.

This work is an initial step into machine learning modeling of the MLD and there are several avenues for continued methodological and oceanographic research. First, the results in this study are regional test cases chosen to reveal how the variability of the MLD impacts the ability of the machine learning methods to learn a functional relationship between the surface variables and the MLD. Future work will expand this regional approach to a global scale. Second, while the probabilistic calibration results suggest that machine learning methods can better estimate the posterior distribution compared to the optimal interpolation approach, the overall calibration is underwhelming. Further research is needed to derive better architectures to better estimate this conditional posterior probability distribution. This research could include weight uncertainty, more sophisticated sampling strategies, covariance regularization, or other neural network architectures. Finally, the research presented in this paper ignored temporal dynamics. We believe that incorporation of the temporal dynamics could help regularize the estimation procedure by coupling observations across time while simultaneously providing useful scientific information about the temporal dynamics of the MLD in relation to the surface variables. In addition to the continued methodological research that follows from this paper, we believe that this methodology can be used to answer scientific oceanographic research questions that require fine resolution gridded MLD estimates.

Acknowledgments

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. Computing and data storage resources, including the Cheyenne supercomputer (doi:10.5065/D6RX99HX), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR. DBW acknowledges useful and motivating discussions with NASA Salinity Science Team colleagues Justin Small, Ivana Cerovecki and Matt Mazloff on contract 80NSSC20K0890.

Code and examples for this project can be found at `https://github.com/NCAR/ml-ocean-bl` and `https://doi.org/10.5281/zenodo.4441098`. Argo-based mixed layer depth data (D. Whitt et al., 2020) can be accessed at `https://doi.org/10.5281/zenodo.4291175`. Preprocessed surface and mixed layer data and model outputs (Foster et al., 2020) can be accessed at `https://www.doi.org/10.5281/zenodo.4421752`. HYCOM data are obtained from `https://www.hycom.org/dataserver/gofs-3pt0/analysis`, experiment GLBu0.08 91.1 for 2014 and experiment GLBu0.08 19.1 for 2012.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016* (pp. 265–283). Retrieved from `https://tensorflow.org.`

Ashukha, A., Lyzhov, A., Molchanov, D., & Vetrov, D. (2020, February). Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. *arXiv e-prints*, arXiv:2002.06470.

Ba, L. J., & Frey, B. (2013). Adaptive dropout for training deep neural networks. In *Advances in Neural Information Processing Systems.*

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015, 07–09 Jul). Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 1613–1622). Lille, France: PMLR. Retrieved from `http://proceedings.mlr.press/v37/blundell15.html`

Bolton, T., & Zanna, L. (2019, Jan). Applications of Deep Learning to Ocean Data Inference and Subgrid Parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. Retrieved from `http://doi.wiley.com/10.1029/2018MS001472` doi: 10.1029/2018MS001472

Brenowitz, N. D., & Bretherton, C. S. (2018, Jun). Prognostic Validation of a Neural Network Unified Physics Parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. Retrieved from `http://doi.wiley.com/10.1029/2018GL078510` doi: 10.1029/2018GL078510

Brooks, S. (2011). *Handbook of Markov chain Monte Carlo.* CRC Press/Taylor & Francis. Retrieved from `https://www.crcpress.com/Handbook-of-Markov-Chain-Monte-Carlo/Brooks-Gelman-Jones-Meng/p/book/9781420079418`

Cabanes, C., Grouazel, A., von Schuckmann, K., Hamon, M., Turpin, V., Coatanoan, C., ... Le Traon, P.-Y. (2013, Jan). The CORA dataset: validation and diagnostics of in-situ ocean temperature and salinity measurements. *Ocean Science*, *9*(1), 1–18. Retrieved from `http://marine.copernicus.eu/services-portfolio/access-to-products/` doi: 10.5194/os-9-1-2013

Caldeira, J., & Nord, B. (2020, April). Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms. *arXiv e-prints*, arXiv:2004.10710.

Caruana, R., Lawrence, S., & Giles, L. (2001, Jan). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems.* Retrieved from http://papers.nips.cc/paper/1895-overfitting-in-neural-nets-backpropagation-conjugate-gradient-and-early-stopping.pdf

Chaudhuri, A. H., Ponte, R. M., & Forget, G. (2016, Apr). Impact of uncertainties in atmospheric boundary conditions on ocean model solutions. *Ocean Modelling*, *100*, 96–108. doi: 10.1016/j.ocemod.2016.02.003

Cintra, R., De Campos Velho, H., Anochi, J., & Cocke, S. (2016, mar). Data assimilation by artificial neural networks for the global FSU atmospheric model: Surface pressure. In *2015 Latin-America Congress on Computational Intelligence, LA-CCI 2015.* Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/LA-CCI.2015.7435937

Cintra, R. S., & Velho, H. F. d. C. (2018, Jul). Data Assimilation by Artificial Neural Networks for an Atmospheric General Circulation Model. In *Advanced applications for artificial neural networks.* Retrieved from http://arxiv.org/abs/1407.4360 doi: 10.5772/intechopen.70791

Cressie, N. A. C. (1993). *Statistics for Spatial Data* (Revised Ed ed.) (No. 1). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved from http://doi.wiley.com/10.1002/9781119115151 doi: 10.1002/9781119115151

Cummings, J. A. (2006, Jan). Operational multivariate ocean data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3583–3604. Retrieved from https://doi.org/10.1256/qj.05.105

Cummings, J. A., & Smedstad, O. M. (2013). Variational data assimilation for the global ocean. In *Data assimilation for atmospheric, oceanic and hydrologic applications (vol. ii)* (Vol. 2, pp. 303–343). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-35088-7_13

Cybenko, G. (1989, Dec). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314. Retrieved from https://link.springer.com/article/10.1007/BF02551274 doi: 10.1007/BF02551274

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., . . . Saurous, R. A. (2017, November). TensorFlow Distributions. *arXiv e-prints*, arXiv:1711.10604.

Dormann, C. F. (2020, Apr). Calibration of probability predictions from machine-learning and statistical models. *Global Ecology and Biogeography*, *29*(4), 760–765. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.13070 doi: 10.1111/geb.13070

Foster, D., Gagne II, D. J., & Whitt, D. (2020, Dec). *Probabilistic Machine Learning Estimation of Ocean Mixed Layer Depth from Dense Satellite and Sparse In-Situ Observations: Preprocessed Satellite and In-situ observation datasets.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.4421752 doi: 10.5281/zenodo.4421752

Fox, D. N., Teague, W. J., Barron, C. N., Carnes, M. R., & Lee, C. M. (2002, Feb). The modular ocean data assimilation system (MODAS). *Journal of Atmospheric and Oceanic Technology*, *19*(2), 240–252. Retrieved from http://journals.ametsoc.org/jtech/article-pdf/19/2/240/3312918/1520-0426 doi: 10.1175/1520-0426(2002)019⟨0240:TMODAS⟩2.0.CO;2

Frankignoul, C., & Hasselmann, K. (1977, Aug). Stochastic climate models, Part II Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, *29*(4), 289–305. Retrieved from http://tellusa.net/index.php/tellusa/article/view/11362 doi: 10.1111/j.2153-3490.1977.tb00740.x

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020, mar). Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz '96 Model. *Journal of Advances in Modeling Earth Systems*, *12*(3). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001896 doi: 10.1029/2019MS001896

Gal, Y. (2016). *Uncertainty in Deep Learning* (Unpublished doctoral dissertation). University of Cambridge.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016* (Vol. 3, pp. 1651–1660). Retrieved from http://yarin.co.

Gal, Y., Hron, J., & Kendall, A. (2017, Dec). Concrete dropout. In *Advances in Neural Information Processing Systems* (pp. 3582–3591).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, third edition.* Taylor & Francis. Retrieved from https://books.google.com/books?id=ZXL6AQAAQBAJ

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018, Jun). Could Machine Learning Break the Convection Parameterization Deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018GL078202 doi: 10.1029/2018GL078202

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press. Retrieved from http://www.deeplearningbook.org

Guinehut, S., Dhomps, A.-L. L., Larnicol, G., & Le Traon, P.-Y. Y. (2012, Oct). High resolution 3-D temperature and salinity fields derived from in situ and satellite observations. *Ocean Science*, *8*(5), 845–857. Retrieved from https://os.copernicus.org/articles/8/845/2012/ doi: 10.5194/os-8-845-2012

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *34th International Conference on Machine Learning, ICML 2017* (Vol. 3, pp. 2130–2143).

Hanawa, K., & Talley, L. D. (2001). Mode waters. *Ocean Circulation and Climate: Observing and Modeling the Global Ocean*, 373–386 (736pp). Retrieved from ftp://bslictb.nerc-bas.ac.uk/jbsall/Papers_CMIP5team/2001Hanawa.pdf

Hernández-Lobato, J. M., & Adams, R. P. (2015). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *32nd International Conference on Machine Learning, ICML 2015* (Vol. 3, pp. 1861–1869).

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012, July). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv e-prints*, arXiv:1207.0580.

Hoffman, M. D., & Blei, D. M. (2015). Structured stochastic variational inference. In *Journal of Machine Learning Research* (Vol. 38, pp. 361–369). Retrieved from http://jmlr.org/papers/v14/hoffman13a.html

Holte, J., Talley, L. D., Gilson, J., & Roemmich, D. (2017, Jun). An Argo mixed layer climatology and database. *Geophysical Research Letters*, *44*(11), 5618–5626. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017GL073426https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL073426https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017GL073426 doi: 10.1002/2017GL073426

Hornik, K., Stinchcombe, M., & White, H. (1989, Jan). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. doi: 10.1016/0893-6080(89)90020-8

Hsieh, W. W., & Tang, B. (1998, Sep). Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography. *Bulletin of the American Meteorological Society*, *79*(9), 1855–1870. Retrieved from http://

journals.ametsoc.org/bams/article-pdf/79/9/1855/3731484/1520-0477 doi: 10.1175/1520-0477(1998)079⟨1855:ANNMTP⟩2.0.CO;2

Irrgang, C., Saynisch-Wagner, J., & Thomas, M. (2020, May). Machine Learning-Based Prediction of Spatiotemporal Uncertainties in Global Wind Velocity Reanalyses. *Journal of Advances in Modeling Earth Systems*, *12*(5). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001876 doi: 10.1029/2019MS001876

Jiang, G. Q., Xu, J., & Wei, J. (2018, Apr). A Deep Learning Algorithm of Neural Network for the Parameterization of Typhoon-Ocean Feedback in Typhoon Forecast Models. *Geophysical Research Letters*, *45*(8), 3706–3716. doi: 10.1002/2018GL077004

Kingma, D. P., & Ba, J. L. (2015, Dec). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.* International Conference on Learning Representations, ICLR. Retrieved from https://arxiv.org/abs/1412.6980v9

Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems* (pp. 2575–2583).

Kingma, D. P., & Welling, M. (2014, Dec). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings.* International Conference on Learning Representations, ICLR. Retrieved from https://arxiv.org/abs/1312.6114v10

Kraus, E. B., & Turner, J. S. (1967, Jan). A one-dimensional model of the seasonal thermocline II. The general theory and its consequences. *Tellus*, *19*(1), 98–106. Retrieved from https://www.tandfonline.com/doi/abs/10.3402/tellusa.v19i1.9753 doi: 10.3402/tellusa.v19i1.9753

Kuleshov, V., Fenner, N., & Ermon, S. (2018, Jul). Accurate uncertainties for deep learning using calibrated regression. In *35th International Conference on Machine Learning, ICML 2018* (Vol. 6, pp. 4369–4377). PMLR. Retrieved from http://proceedings.mlr.press/v80/kuleshov18a.html

Labach, A., Salehinejad, H., & Valaee, S. (2019, April). Survey of Dropout Methods for Deep Neural Networks. *arXiv e-prints*, arXiv:1904.13310.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017, Dec). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (pp. 6403–6414).

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016, Jan). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. doi: 10.1016/j.gsf.2015.07.003

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993, Jan). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867. doi: 10.1016/S0893-6080(05)80131-5

Maeda, S.-i. (2014, December). A Bayesian encourages dropout. *arXiv e-prints*, arXiv:1412.7003.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., ... Williams, J. K. (2017, Oct). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. doi: 10.1175/BAMS-D-16-0123.1

Melnichenko, O., Hacker, P., Maximenko, N., Lagerloef, G., & Potemra, J. (2016, Jan). Optimum interpolation analysis of Aquarius sea surface salinity. *Journal of Geophysical Research: Oceans*, *121*(1), 602–615. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/2015JC011343 doi: 10.1002/2015JC011343

Monteleoni, C., Schmidt, G. A., & McQuade, S. (2013, Sep). Climate informatics:

Accelerating discovering in climate science with machine learning. *Computing in Science and Engineering*, *15*(5), 32–40. doi: 10.1109/MCSE.2013.50
Neal, R. (1996). Bayesian Learning for Neural Networks. *Lecture Notes in Statistics*, *1*(118).
Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., & Tran, D. (2019, April). Measuring Calibration in Deep Learning. *arXiv e-prints*, arXiv:1904.01685.
O'Gorman, P. A., & Dwyer, J. G. (2018). Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. doi: 10.1029/2018MS001351
Ouali, D., Chebana, F., & Ouarda, T. B. (2017, Jun). Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. *Journal of Advances in Modeling Earth Systems*, *9*(2), 1292–1306. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000830 doi: 10.1002/2016MS000830
Paisley, J., Blei, D. M., & Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* (Vol. 2, pp. 1367–1374).
Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018, Jan). Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters*, *120*(2), 024102. Retrieved from https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.024102 doi: 10.1103/PhysRevLett.120.024102
Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press. Retrieved from https://mitpress.mit.edu/books/gaussian-processes-machine-learning
Rasp, S., Pritchard, M. S., & Gentine, P. (2018, Sep). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9684–9689. doi: 10.1073/pnas.1810286115
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019, Feb). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. Retrieved from https://www.nature.com/articles/s41586-019-0912-1 doi: 10.1038/s41586-019-0912-1
Remote Sensing Systems. (2017). *GHRSST Level 4 MW_OI Global Foundation Sea Surface Temperature analysis version 5.0 from REMSS.* NASA Physical Oceanography DAAC. Retrieved from https://podaac.jpl.nasa.gov/dataset/MW_OI-REMSS-L4-GLOB-v5.0 doi: https://doi.org/10.5067/GHMWO-4FR05
Rezende, D. J., Mohamed, S., & Wierstra, D. (2014, Jan). Stochastic backpropagation and approximate inference in deep generative models. In *31st International Conference on Machine Learning, ICML 2014* (Vol. 4, pp. 3057–3070). Retrieved from http://proceedings.mlr.press/v32/rezende14.html
Roemmich, D., & Gilson, J. (2009, Aug). The 2004-2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in Oceanography*, *82*(2), 81–100. doi: 10.1016/j.pocean.2009.03.004
Rosenblatt, F. (1958, Nov). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. doi: 10.1037/h0042519
Ruder, S. (2016, September). An overview of gradient descent optimization algorithms. *arXiv e-prints*, arXiv:1609.04747.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representa-

tions by back-propagating errors. *Nature*, *323*(6088), 533–536. Retrieved from https://www.nature.com/articles/323533a0 doi: 10.1038/323533a0

Schmidtko, S., Johnson, G. C., & Lyman, J. M. (2013, Apr). MIMOC: A global monthly isopycnal upper-ocean climatology with mixed layers. *Journal of Geophysical Research: Oceans*, *118*(4), 1658–1672. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/jgrc.20122 doi: 10.1002/jgrc.20122

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stommel, H. (1979, Jul). Determination of water mass properties of water pumped down from the Ekman layer to the geostrophic flow below. *Proceedings of the National Academy of Sciences*, *76*(7), 3051–3055. Retrieved from https://www.pnas.org/content/76/7/3051https://www.pnas.org/content/76/7/3051.abstract doi: 10.1073/pnas.76.7.3051

Ukkonen, P., & Mäkelä, A. (2019, Jun). Evaluation of Machine Learning Classifiers for Predicting Deep Convection. *Journal of Advances in Modeling Earth Systems*, *11*(6), 1784–1802. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001561 doi: 10.1029/2018MS001561

Valler, V., Franke, J., & Brönnimann, S. (2019). Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation. *Climate of the Past*, *15*(4), 1427–1441. Retrieved from https://doi.org/10.5194/cp-15-1427-2019 doi: 10.5194/cp-15-1427-2019

Wahle, K., Staneva, J., & Guenther, H. (2015, Dec). Data assimilation of ocean wind waves using Neural Networks: A case study for the German Bight. *Ocean Modelling*, *96*, 117–125. doi: 10.1016/j.ocemod.2015.07.007

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011* (pp. 681–688).

Weyn, J. A., Durran, D. R., & Caruana, R. (2019, Aug). Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001705 doi: 10.1029/2019MS001705

Whitt, D., Nicholson, S., & Carranza, M. (2020, November). *Argo-based ocean surface mixed layer depths using the buoyancy gradient definition of Whitt Nicholson and Carranza (2019).* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.4291175 doi: 10.5281/zenodo.4291175

Whitt, D. B., Nicholson, S. A., & Carranza, M. M. (2019, Dec). Global Impacts of Subseasonal (<60 Day) Wind Variability on Ocean Surface Stress, Buoyancy Flux, and Mixed Layer Depth. *Journal of Geophysical Research: Oceans*, *124*(12), 8798–8831. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015166 doi: 10.1029/2019JC015166

Zlotnicki, V., Qu, Z., & Willis, J. (2019). *MEaSUREs Gridded Sea Surface Height Anomalies Version 1812.* NASA Physical Oceanography DAAC. Retrieved from https://podaac.jpl.nasa.gov/dataset/SEA_SURFACE_HEIGHT_ALT_GRIDS_L4_2SATS_5DAY_6THDEG_V_JPL1812 doi: 10.5067/SLREF-CDRV2